1990

11. The Assessment of Teacher Assessment: Concluding Thoughts and Some Lingering Questions

James V. Mitchell Jr.
University of Nebraska-Lincoln

11

The Assessment of Teacher Assessment: Concluding Thoughts and Some Lingering Questions

James V. Mitchell, Jr.
University of Nebraska-Lincoln

In a recent update on teacher testing practices across the United States, Rudner (1988) reported that 44 states have developed teacher-certification-testing programs, with 26 states currently testing prospective teachers as a certification requirement and another 18 states scheduled to implement such programs in the near future. It is obvious that teacher testing has become a very extended endeavor. It has also stimulated extended debate.

It was in acknowledgment of the importance of this extended teacher testing and associated debate that the Advisory Committee of the Buros Institute of Mental Measurements decided to devote its 1987 annual symposium to the topic of teacher assessment. As we developed the plans for this symposium, we tried to keep in mind two principles to guide our thinking and planning: (a) our treatment of teacher assessment was not to be narrowly conceived and focused on a singular aspect of teacher assessment (e.g., assessment for certification), but rather was to address the larger measurement and implementation issues that were generic to many or all teacher assessment settings; and (b) we hoped that we could avoid the mere rehashing of old issues and instead effectively advance thinking about teacher assessment in ways that, in

John Dewey's words, would represent "a level deeper and more inclusive than is represented by the ideas and practices of the contending parties" (Dewey, 1949, p. v). We hope that we have accomplished that, at least to some degree, both in the symposium and now with the book. The purpose of this concluding chapter is to work within this context to highlight and compare some of the salient thoughts of the several contributors, to reflect on their meaning and implications, and to point out some of the issues that remain. Each contributor is considered in turn, with summary comment to follow about their combined contributions. W. James Popham, the keynote speaker, is considered first.

When we first asked Jim Popham to present the keynote address at the symposium, we had in mind both his extended and important contributions to this area and the fact that this experience would qualify him admirably for addressing the questions implied by the topic we had tentatively suggested: "Teacher Assessment: Why and for What Purpose?" When he accepted the invitation, Popham asked whether he could "spice up" the title, and the final result was a "spicing up" of both title and topic by focusing on an issue that he felt had very potent implications for the future of teacher assessment: "Face Validity: Siren Song for Teacher Testers."

Popham's contention that we are being lured away from more important concerns by becoming preoccupied with face validity considerations is an important and timely one for many participants in the teacher-testing enterprise, particularly those who are not as indoctrinated with the holy trinity of validity classifications as most measurement people are. But measurement people are only a small contingent in the teacher-testing arena, and the lure of face-validity considerations over the more important consideration of the validity of score-based inferences is but another example of the miscommunication, differing (and sometimes unknowledgeable) expectations of different groups, and downright wish fulfillment that often seems rampant whenever the issue of teacher testing arises. The "quick fix" mentality often found in the public, legislators, governmental agencies, and even in some educators creates a setting where the siren song of face validity becomes irresistibly appealing. Popham is to be congratulated for warning us of the risk.

I am almost totally in accord with the major points made by Popham, and this should be kept in mind in the following discussion. However, there were some issues that were raised in my mind

that did not necessarily lessen the effect of Popham's arguments but were stimulated by the major directions that his arguments took. If these are side issues, they are important side issues, and they are an interesting example of how a focus on one particular issue can raise other issues for which the answers sought are important in their own right as well as for their contribution to the understanding of the original issue.

The first issue relates to Popham's definition of face validity, a definition that I believe most of us would find acceptable: "*Face validity constitutes the perceived legitimacy of a test for the use to which it is being put.*" As I read that definition I was struck by the extent to which "perceived legitimacy" plays a role not only in the face-validity setting but also in the content-validity exercises that are so much a part of the local validation effort for teacher tests like the National Teacher Examinations (NTE). It is sometimes hard to determine why "perceived legitimacy" is accorded so much more professional approval in the case of content validity than it is for face validity. For the NTE, for example, a typical content validity exercise would have a college-based panel address the question of the content appropriateness of each test item by asking each panelist whether 90% of the applicants for entry-level certification have had the opportunity to acquire the knowledge or academic skills being tested; another panel, in this case a school-based panel, would address the question of the job relatedness of each test item by asking each panelist how important the knowledge or skill was for the beginning teacher in general. If this isn't a "perceived legitimacy" question, I don't know what is. There are differences, of course, but are the differences critical? In the case of the NTE panels, for example, the panelists are supposed to be either experts or very knowledgeable people who have direct personal experience with the content or job that defines the judgment setting. Face-validity judgments usually refer to judgments by less qualified or knowledgeable people. Another difference is what is judged. The NTE panels judge either content relatedness with teacher-training curricula or relatedness to the job of teaching. Face-validity considerations involve judgments about whether the test or test items look appropriate for the testing of teachers. But they are both "perceived legitimacy" judgments with all the human frailties usually associated with such judgments.

The greater respect and status accorded to content validity exercises of the kind employed for the NTE must reside in the knowledgeability ascribed to the judges and the relatively systematic methods used in arriving at the judgments. This is in contrast to

the naivete or lack of knowledgeability presumed to be present in the face validity situation and the unsystematic and impressionistic judgments that are supposed to characterize that situation. The drawing of these contrasts is forced, however, in the light of the indeterminacy and confusion that often accompany these NTE content validity exercises. In contrast to face validity, content validity is supposed to involve a kind of focused rationality, but in practice it can be noticeably short in both focus and rationality. Face validity is certainly the ogre that Jim Popham says it is, but the content-validity measures that are often offered as the antidote have much more in common with the inadequacies of face-validity judgments than we are commonly willing to recognize. "Perceived legitimacy" is a siren song for teacher testers wherever it appears, and its ill effects can be felt in the presumedly more antiseptic environment of the content-validity exercise as well as the nasty context of face-validity perceptions. Both involve perception and judgment, and while we are casting out the one we should recognize that the other, pure and white because of its presence in the holy trinity of measurement, evokes the same kind of cognitive processes and is susceptible, hopefully but not assuredly in lesser degree, to the same kinds of bad habits.

The second issue that was raised in my mind as I read the Popham chapter was again not an issue prompted by disagreement, but rather an issue stimulated by the development of the argument. I took special note of Popham's contention that: "Typically, in the case of teacher tests, we administer tests so that we can make inferences about how a teacher is apt to behave in an instructional setting." This seems to connote a confidence in the predictive efficacy of teacher tests that is not shared by all members of the measurement community. This confidence is shown again in a concluding statement: "If, however, an increase in face validity causes a decrease in the validity of score-based inferences, then efforts to enhance face validity should be foregone." If something can cause a decrease in the validity of score-based inferences, the clear implication is that there was some legitimate validity to begin with. An in-house review of the validity of the earlier National Teacher Examinations revealed a median correlation of .11 for seven studies that involved the correlation of a weighted total score for the Common Examinations and ratings by supervisors and principals during the 1st year of teaching (Quirk, Witten, & Weinberg, 1973). This does not inspire confidence. Yet even within the group of authors contributing to this volume there are substantial differences about the kinds of validity evidence required and

whether acceptable validity levels have been achieved. Popham and Mehrens, for example, are optimists on these issues, and Madaus and Mitchell take a much more pessimistic view.

Regardless of one's requirements or interpretation of the data, however, I find it somewhat difficult to conceive of many situations where you would actually be required to trade off the validity of score-based inferences for face validity, especially if in the former case we were referring to the criterion-related validity suggested by Popham's statement that we administer these tests "so that we can make inferences about how a teacher is apt to behave in the instructional setting." I can think of possible situations where we might have to trade off someone's judgment about the curricular validity or job relatedness of test content for someone else's judgment about whether test content seems to have perceived legitimacy for the job of teaching. If I stated it in its most outrageous terms, I would say that the choice is sometimes between face validity and no validity or face validity and pretend validity. In that kind of situation I'm not sure that I would even bother to make the choice; the choice is not worth making and the test is not worth giving.

Popham's major point, however, is well worth making and is cogently argued: Face validity can be a snare and a delusion. Unfortunately, content validity is sometimes susceptible to very similar ills.

Edward Haertel's chapter on "Teacher Performance Assessments: A New kind of Teacher Examination" is a very useful status report on the current work of the Teacher Assessment Project (TAP), sponsored by the Carnegie Corporation and under the direction of Lee S. Shulman, from the point of view of an active participant in the project. In a report entitled *A Nation Prepared: Teachers for the 21st Century* (Carnegie Forum on Education and the Economy, 1986) the Carnegie Foundation Task Force on Teaching as a Profession recommended the creation of "a National Board for Professional Teaching Standards, organized with a regional and state membership structure, to establish high standards for what teachers need to know and be able to do, and to certify teachers who meet that standard" (p. 55). The measurement problems inherent in such an effort are immense, of course, and the Teacher Assessment Project has accepted the heady task of developing the teacher performance assessments that might serve as the basis for board certification.

Haertel's discussion of the TAP project must be in the nature of

unfinished business, because the project has been in existence for only a short period of time. The nature of some of the challenges to come, however, may be clearly grasped from descriptions of steps taken to date. One particularly important statement, in my judgment, is the following:

> By design, nearly all of the questions posed in the various exercises have several correct answers. In scoring, it is necessary to recognize the validity of alternative instructional approaches while maintaining distinctions among different degrees of response quality (p. 23).

The immensity and inherent difficulties of the task of developing valid scoring procedures for the TAP structured performance assessments are here cast in bold relief. It is admirable for teacher-assessment procedures to have the flexibility that acknowledges the many alternative patterns that may constitute effective teaching, but the determination of which of many alternative patterns constitute "correct" answers and have the requisite "response quality" will be certain to magnify appreciably the problems of establishing the validity of the assessment. One of the first problems will be a classificatory problem: classification in terms of the essential features of the phenomenon being observed, and classification in terms of the quality of response observed. The TAP program has chosen initially to structure the first classification issue in terms of five "scoring dimensions" described as "content-specific pedagogy," "subject matter knowledge," "professional responsibility," "class organization and management," and "pedagogy, sensitivity, and responsiveness to students." "Professional responsibility" and "pedagogy, sensitivity, and responsiveness to students" are examples of such broad and ambiguous categories that one wonders about the definitiveness of any inclusion criteria that could be developed and the reliability of the assignments to such categories. The reliability of assignments to these categories raises another problem that Haertel also acknowledges in his own discussion: Will these categories have the convergent and discriminant validity that is required of them, or will cross-dimension correlations be so high as to negate their hoped-for utility?

The classification or scoring of response quality is not without its problems, either. The attempt to identify discrete, scorable elements of a teaching performance and then to combine these scorable elements "following a more or less explicit rule" is a procedure having evident heuristic value but also one that cries for reliability and validity evidence that provides some ultimate justification for

the ad hoc nature of the approach. Haertel rightly warns us about premature insistence on validity evidence, but validity evidence is critical and cannot be long postponed if the people most immediately involved are to have confidence in the procedures developed. The alternative scoring procedure described, "holistic" scoring, involves a matching of performance elements to descriptions of previously rated prototype performances. This matching process seems to proceed in terms of relatively inexplicit criteria as well, and once again its justification can only come from sound evidence of its reliability and validity.

All of these problems are doubtlessly well known to Haertel, the TAP director, and their collaborators. It is a brave effort and a very necessary one, and it is imperative that the effort be supported and the problems addressed. There are two major dangers that concern me. The first is that the validity and scoring problems will be so time consuming that, eventually, compromises will have to be made and there will be a retreat to face validity justification of the type Jim Popham decried. Because face validity could certainly be ascribed to these exercises, the temptation would always be there. The second danger is that the procedures themselves will be found to be so time consuming that those who inherit them will have neither the time, patience, nor professional expertise to apply them as well as their developers. Such conditions would again have immense consequences for validity.

Donald Medley's chapter on "Improving Teaching Through the Assessment Process" is a tribute to systematic thinking and the scientific method. His model of the teaching-learning process helps to avoid much of the conceptual confusion that can occur in the teacher-assessment area long before any of the measurement issues are addressed. His choice of measurement techniques then follows naturally from the model chosen, and the reader can appreciate the final result in the context of its conceptual underpinnings. His model also acknowledges the complexity of the teacher-assessment problem and provides a manageable structuring of that complexity to facilitate understanding about where one might most effectively enter the system and how one might most effectively take advantage of its measurement implications. Scholars can also be realists, and it is this happy combination that makes Medley's contribution so worthy of careful thought. He more than proves the case for his contention that "there is a better way to assess competencies than the conventional tests presently used."

The last section of Medley's chapter is devoted to assessment

procedures that meet the requirements of his model and focus on the "skills . . . a teacher needs to do well during interactive teaching," which he perceives as having little in common with the skills a student needs to do well on a conventional multiple-choice test of professional knowledge. The following comments are directed to the two assessment procedures discussed in this part of the chapter: the simulation exercises and the instrumentation used to assess teacher competence in the Beginning Teacher Assistance Program (BTAP).

The two simulation exercises illustrated are doubtlessly far more related to the problems a teacher in the classroom would face than the typical item in a typical teacher assessment test. In this sense Medley has done what he set out to do. Both exercises represent a face validity that has inherent appeal; at the same time both conjure up some of Popham's concerns about the issues that may lie beyond initial impressions of face validity. Verisimilitude in relation to actual classroom problems is certainly a strong characteristic of these simulation exercises, but that verisimilitude does not necessarily guarantee that the responses will better predict what that same person would do in the actual classroom situation with the same problem. There are at least four concerns that arise in relation to simulation exercises of the type described: (a) demand characteristics of the setting; (b) fakability, (c) the affective components of most problem situations in teaching, and (d) the determination of an acceptable definition of what constitutes "professional knowledge."

Demand characteristics are different for different exercises, and the demand characteristics of the first exercise (the cheating episode) are probably the strongest of the two simulation exercises. In view of the fact that the two participants in the cheating episode are described as normally well behaved and even docile, is there much doubt that the more severe punishments for cheating would be regarded as unacceptable by the powers that be? Because the cues for what is wanted are likely to be stronger and more evident than the cues for what the respondent would *actually* do in the situation (which may be unpredictable even to the respondent), and because the motivation would also be correspondingly stronger for these demand cues, the "wrong" responses can be rather easily eliminated for the wrong reasons. The influence can be unconscious in nature, but in most cases it will probably be conscious and will then demonstrate what we usually refer to as the "fakability" of the item. One can always take the position that the exercise represents professional knowledge, not actual profes-

sional performance in the classroom, but in that event the exercise is not better than an item similarly unpredictive of actual performance that is totally without verisimilitude. Neither way can we predict accurately what the person is going to do in the classroom.

Another concern is that exercise verisimilitude is incomplete at best if it only reflects the cognitive components of a problem situation. Many problem-solving situations in the classroom have strong affective overtones caused both by the teacher's own affective needs and self-concept and by the nature of the teacher's previous and present interactions with pupils in the classroom. It is probably impossible to reflect these affective variables in any teacher-assessment procedure except an actual classroom observation, but their existence does make prediction difficult even from the most realistic of problem depictions. One can again retreat to the position that it is professional knowledge that is being assessed, not actual professional performance, but again one can question what gains have really been made over conventional methods by the effort to create verisimilitude—not predictive gains, certainly.

Still another concern is the problem of defining what constitutes professional knowledge, the very gist of what is being assessed. Many of those who are cognizant of the literature on the prediction of teaching effectiveness would probably conclude that well-verified empirical results in this area are few and far between and the evidence uncertain and inconclusive. It is interesting to note that the first simulation exercise was taken from a self-instructional package developed for in-service teacher education and that the keyed answers were those consistent with the "recommendations" given by those who developed the packets. One can wonder whether this kind of "knowledge" is deserving of the term *professional knowledge* or whether it is a combination of common sense, good judgment about the probable consequences of actions, and personal and professional values. Even if it did not meet the criteria of empirically verified knowledge, perhaps it could be excused for that if it adequately predicted performance in the classroom. But there is no evidence that it does that, either; if we are realistic we have to challenge both its legitimacy as "knowledge" and its predictive efficacy for actual classroom performance.

Medley acknowledges that "interactive performance skills," unlike professional knowledge, cannot be assessed adequately by simulation. The very different kinds of interactional settings that exist between students and teacher cannot be accurately reflected in the typical simulation, and other methods must be sought. For

this he advocates the measurement-based teacher evaluation that was implemented in the Virginia Beginning Teacher Assistance Program (BTAP). There is little doubt that the Virginia program constitutes one of the most credible and creditable programs around, and it is much advanced over most of its competitors. It is not the simulation of setting and the simulation of performance choice; it is the actual setting and the actual choice and execution of performance options. What has been done to develop and implement the program is impressive. The fact that certain questions can still be raised about the program is not so much a reflection of program shortcomings as it is a reflection of the complexity of the problems that can bedevil any effort to assess teacher performance.

Although the BTAP program reduces some of the problem of generalizability that occurs with simulation, some of the problems remain, even if in somewhat different form. Demand characteristics are still very much in evidence. With the simulation exercises it is the problem of showing that you know and can choose the correct professional knowledge. With the BTAP program someone else has chosen the professional knowledge, in this case the 70 research-based categories of teacher behavior labeled *indicators of competence,* and you have to demonstrate the behavior required. But whether you choose on demand or act on demand, there is still the very real problem of what you are really going to do when the demands are removed. If the problem setting is not effective in predicting that future behavior, and if fakability is still a serious issue, both the predictive efficacy of the program is in question, and the potential of the program to improve future teaching behavior can be seriously doubted.

Because teachers are required to demonstrate "indicators of competence" (e.g., ending a unit with a summary or review) in an actual classroom setting, the BTAP program does much more than the simulation to include both the affective and cognitive elements that together produce the climate of a real teaching situation. This is a decided plus. The setting is no doubt influenced, however, by the likelihood that the teacher is showing off his or her very best behavior, and the affective requirements and responses for this setting may be quite different from what occurs when the observer leaves the classroom.

The knowledge base for the "indicators of competence" is also at issue. The 70 categories of teacher behavior that served as the knowledge base were identified from the research literature on teacher effectiveness. The strength and relevance of these research

findings are a matter of professional judgment, and that judgment is likely to be varied. The weights to be assigned these indicators in anyone's implicit set of judgments about what constitutes effective teaching is also open to question; it certainly seems that in any implicit *or* empirical set of weightings for defining effective teaching, "making interrelationships among parts of the lesson clear to learners" should be assigned greater importance than "beginning the lesson or unit with a statement of purpose." What is one person's knowledge base seems to be another person's morass of inconclusiveness. In such a setting professional value judgments seem to play a large role.

The issues of scoring procedures and passing scores also loom large in an assessment undertaking of this nature. According to Medley, "A temporary scoring key was constructed for each of the 14 competencies by first identifying a set of events that reflected the indicators that defined that competency, and summing the standard scores . . . in each record" (p. 69). Subsequent revisions of the keys were undertaken to maximize coefficient alpha. To be included an indicator had to be perceived as supported by research on teaching, and once included its weight appeared to be equal to weights of all other indicators included. The pass score was based on an estimate by principals of what percentage of teachers in the state possessed that competency. This is another judgment game that does not have very precise rules. What is important to recognize here is that the actual observing and recording procedures in this system are low-inference procedures; where the high inference occurs is in selection of indicators, the scoring procedures, and the setting of the passing scores. High-inference procedures, wherever they occur, need constant study and verification.

Because the complexities are so great, any teacher-assessment program will have its stronger points and its weaker points. It is far easier to critique than to create. Donald Medley has created a conceptual scaffolding and a teacher-assessment program that demonstrate remarkable improvements over earlier state-of-the-art efforts, and his work has resulted in major contributions that have advanced and will advance teacher assessment for some time to come.

William Mehrens' "Assessing the Quality of Teacher-Assessment Tests" is an extremely comprehensive and useful compilation of facts and insights about the development of teacher-assessment tests and methods of assessing and assuring their quality. Of particular interest is the extensive treatment of validity considera-

tions in the development of teacher-assessment tests. Readers of this volume encounter very different judgments about the adequacy of most current efforts to establish validity for teacher-assessment instruments; Mehrens probably represents the most optimistic end of the continuum and George Madaus represents the most pessimistic end. Perhaps this author ought to indicate his predilections before commenting further about the Mehrens' chapter, for I must admit to a more pessimistic view of present methods of establishing validity evidence for teacher-assessment tests and also the quality of validity evidence so produced. With this as context I offer the following as a basis for discussion.

Mehrens spends a great deal of time and effort in the discussion of methods of establishing *content* validity for teacher-licensure tests, and his faith in the procedures and results of these content-validation efforts is admittedly much greater than mine. His discussion of the development of a list of competencies, the analysis of job requirements, the development of test specifications, and the development and validation of items is thorough, thoughtful, and stimulating. But as I read this discussion I could not help being impressed by two statements that summarized the basic weakness of the foundation of the entire structure. In speaking of the test development procedure just described, it is stated that "It [the test] will be assessing those competencies that experts in the field *thought necessary* [italics added] for beginning professionals to have in order to protect the public" (p. 99). Then later, in a discussion of criterion-related validity, it is indicated that "there is no clear definition of what it means to be an effective teacher" (p. 102), with a reference to a paper by Webb (1983).

The appeal to authority, to "experts in the field," whether they be practicing elementary or secondary teachers or university professors of education, is not one to inspire confidence, especially in view of the fact that there is so little agreement among experts or anybody else on "what it means to be an effective teacher." The latter has just as many implications for content validity as for criterion-related validity. Whether one is looking for the ideal empirical criterion or the ideal of teaching effectiveness as fashioned by several disagreeing "experts," the goal is just as unattainable. The "thought necessary" criterion, stripped of its verbal superstructure, is nothing more than simple opinion, expert or not, and it is simple opinion based on nonexistent or at least unimposing scientific findings. Sometimes it can be little more than ideology or value judgment. Furthermore, there is little doubt that often the "experts" who are asked to apply the "thought necessary" or sim-

ilar criteria are uncertain and frustrated about their task. Whether they actually feel "expert" in either their final judgment or the certainty with which that judgment is held is open to serious question.

The "thought necessary" criterion applied by "experts in the field" seems particularly worrisome in the context of Webb's (1983) painfully evident contention that there is no clear definition of what it means to be an effective teacher. The paragraph in which she expresses that contention is a thought-provoking one that deserves to be quoted in its entirety:

> Although no one would question the importance of good teaching to the provision of good education, the appraisal of teacher performance has presented numerous and nettlesome problems. One major problem inherent in teacher evaluation is that there is no clear definition of what characterizes an effective teacher or constitutes effective teaching and, consequently, no definitive measures to be used for teacher evaluation. Any evaluation process is essentially a comparison of desired outcomes with actual outcomes. If the situation exists where not only the results but in many cases the desired outcomes are in question, then the task of evaluation becomes extremely difficult. (Webb, 1983, p. 69)

Although Webb seems to be talking about teacher assessment in general, she is also talking about evaluation involving minimum competency testing, and her comments apply with equal force to assessment for licensure. If there is no clear and agreed-upon definition of what it means to be an effective teacher, the specific competencies of an effective teacher will be difficult to define or agree upon, and the further difficult task of defining what *minimum* levels of competencies should be for licensure or other purposes becomes an unstructured and confusing enterprise. And it has often been just that. Through compromise and adjustment, teacher-assessment instruments do get constructed, but the gap between the ideal of content validity and the actuality of practice typically makes the final product very vulnerable to challenge.

Mehrens' discussion of the establishment of validity evidence leaves no doubt that it is content validity evidence that should shoulder all the burden. He cites several authorities who have argued that it is both "unfeasible and inappropriate to expect criterion related validity of a licensure examination." He quotes the *Standards for Educational and Psychological Testing* (1985) to the same effect. He cautions that when a rating is used as a criterion, and a test as a predictor, it is difficult to determine whether a

failure of the test to predict the rating is the fault of the test or of the rating.

In view of Mehrens' conviction of the unfeasibility of collecting criterion-related validity evidence for licensure tests, I am a little surprised that at the end of his discussion he indicates that:

> It does not follow from all of the preceding statements that it is inappropriate to attempt to find out what, if any, correlates of teacher-licensure tests exist. Although correlational data are somewhat sparse, they are consonant with the logical inference that knowledge about teaching and the subject matter being taught (competence) should be related to both performance and effectiveness in teaching. (p. 104)

Mehrens then reports data he apparently feels is consistent with this statement. Two of the studies he cites involved the correlation of NTE scores with aptitude tests. In what way does teaching performance or effectiveness play a role in either one of these? Is NTE now the criterion instead of the predictor? High correlations between general ability and achievement tests have been recognized for a long time, but the relationship says nothing about how knowledge about teaching might be related to actual teaching performance or effectiveness. Mehrens cites another study (Piper & O'Sullivan, 1981) as reporting a correlation of .43 between the NTE Common Examinations scores and a supervisor's rating on a Performance Evaluation Instrument. Actually, the correlation of .43 was between the NTE Elementary Area examination and the Performance Evaluation Instrument. The study was a half-page brief research report with only 32 subjects, and there were aspects of the study that were puzzling and required additional explanation. Overall, it may be better not to depart at all from one's contention that criterion-related validity evidence is unfeasible than to place much reliance on data of this type.

The Mehrens' chapter stimulated a great deal of thought on my part, and that is a tribute to its author. There are times when I think we should change our direction completely. Predicting teacher performance or effectiveness (even for licensure purposes) may be a pretty hopeless task. There may be a problem of what, for want of better terminology, I refer to as a "shifting criterion": There may be 1001 ways of being an effective teacher, and 1001 ways of being an ineffective teacher. This only acknowledges the complexity of what goes on in teaching. There are other ways to satisfy the public's concern about teachers. Why not start with the

basic proposition that it simply isn't good to have an ignoramus in front of an elementary or secondary classroom? Don't suggest or imply that this has anything to do with effective teaching, because we really don't know the point at which it does; it is simply desirable, for modeling and social learning purposes alone, to have a well-educated person in front of either elementary or secondary students. Forget your list of competencies or your job analyses; they sometimes seem to contribute more to problem confusion than problem solution. Require instead that prospective teachers pass a general education examination that attests to their achievement in reading, writing, speaking, mathematics, reasoning, and knowledge of the culture. Require that this examination be passed before the person can be admitted to a teacher-education program. Do not pretend that this will ensure effective teachers, and strongly disabuse the public of any such thoughts. Don't do that which we may not be able to do; do that which should be done for its own sake, and avoid the tortuous route of collecting content-validity evidence that may have very limited or no meaning and even less predictive efficacy. And let the higher-education subject areas, in cooperation with teacher-education specialists, determine whether a prospective teacher has sufficient knowledge to teach a given subject at the secondary level. If all this seems a little iconoclastic, it should be attributed to this author's continuing frustration with the validity problems of teacher-assessment tests and the conviction that a better direction must be sought.

For those who feel that content-validity exercises, as they are now or as they will presumably be improved, are the royal road to better teacher assessment (or licensure) tests, William Mehrens points the way with his usual thoroughness and illuminating insights. For those like me who are experiencing doubts and frustrations, Mehrens' contributions serve masterfully to stimulate careful thought about where we are, where we are going, and what the alternatives might be.

Linda Darling-Hammond's chapter on "Teacher Evaluation in the Organizational Context" serves an extremely useful function for this volume on assessment of teaching. When the symposium and book were being planned, it was felt that the influence of organizational context on the nature, processes, and results of teacher assessment was so profound that we had to include one chapter that would emphasize the importance of context and shed some useful light on the specific processes by which these contextual influences affect the nature and implementation of teacher

assessment. Darling-Hammond's chapter does that extremely well. Her focus is not at the state level and on licensure examinations but rather on evaluation activity as it occurs at the teacher, school, and system level. She forces us into some healthy reality testing by delineating convincingly the several influences of the organizational context and demonstrating the consequences if these are poorly understood or ignored. The chapter stands by itself, and thus my comments will be restricted.

When all the organizational context variables are brought to bear, as they are in this chapter, one can develop a much more profound grasp of why teaching presents so many difficult problems for evaluation. The descriptions of the various "Conceptions of Teaching Work," with teaching conceived alternatively as labor, craft, profession, or art, underscore the very different purposes and procedures that would govern the evaluation of teaching so differently conceived. The cited experience of the Beginning Teacher Evaluation Study conducted for California's Commission for Teacher Preparation and Licensing is interesting in relation to the earlier discussion of Mehrens' chapter; they concluded that their findings suggested "that the legal requirement for a license probably cannot be well stated in precise behavioral terms" (Bush, 1979, p. 15). Interesting also is the evidence given that the effort to specify specific teaching behaviors related to increased student achievement can often result in two- or three-way interactions that are difficult to translate into rules of practice. The generalizability of such interactions for classroom practice is thus severely constrained. Furthermore, any relationships found with achievement are often curvilinear, which provides additional limits on generalizability. She reports that "Research on nonteaching variables in the educational environment indicates that many factors other than teaching behaviors have profound effects on student learning" (p. 146), and later quotes approvingly from Doyle's (1979) statement that such an ecological approach, which acknowledges the influence of important nonteaching variables on achievement, "would seem to call into question the very possibility of achieving a substantial number of highly generalizable statements about teaching effectiveness" (Doyle, 1979, pp. 203–204). Thus performance-based teacher-evaluation models, based on the presumption that there are generalizable rules for teaching behavior that will lead to increased student achievement, are procrustean models that too often fail to acknowledge the contextual complexity of teaching and fall short as a result. Predetermined approaches to teaching, and their associated predetermined approaches to the

evaluation of teaching, fail to acknowledge the many student, classroom, and school variables that the effective teacher must react to in the decision making that will ultimately shape his or her teaching behavior. For once one senses in this discussion a forthright recognition of the true complexity of what we are up against. It is refreshing.

It is equally refreshing to benefit from Darling-Hammond's discussion of the logistic, financial, and political realities that have so much impact on the usefulness of an evaluation program. She reports Knapp's (1982) contention that in actual practice schools follow the lines of least resistance and evaluate "aspects of teachers and teaching in more vague terms so as to simultaneously satisfy diverse constituencies." This is a humbling statement, but the satisfaction of these diverse constituencies is doubtlessly a very potent political reality and an accurate portrayal of the typical setting. She quotes a very interesting statement from Knapp (1982) that includes a sentence to the effect that "Value choices are nowhere more clearly at issue than in decisions about the aspects of the teacher and teaching to be evaluated" (p. 4). This acknowledgement of the importance of *values* in teacher evaluation is a very critical consideration. If research on teacher effects does not provide as strong a foundation for teacher evaluation as we should like, and if logistic, financial, and political influences are prepotent in the typical setting anyway, it is likely to be *values* that will have as much or more influence than anything else on what is finally developed for teacher-evaluation purposes. The process of "satisfying diverse constituencies" will eventually result in an averaging process that represents the "lines of least resistance," and the value directions emerging will somehow get embedded in the final teacher-assessment program. The final program may not reflect that imperfectly emerging value system as well as it should or as well as most people might have hoped, but that may be a function of the vagueness already referred to and the difficult leap from value perspective to instrumentation. It is realistically useful, in my judgment, to recognize the important role of value choice in the development of teacher-evaluation programs; if you don't know what an effective teacher is, construct an effective teacher from your value repertoire and try to embed it in your evaluation system. This may sound cynical, but it is only meant to prevent us from kidding ourselves about what most teacher-evaluation systems actually are. Any item in any teacher-evaluation instrument I have seen reflects a value about a characteristic a good teacher is supposed to have. We may hope that such items are

based on what little research evidence we have, but they are most likely influenced greatly by the political realities Darling-Hammond discusses so trenchantly. Those political realities are ignored only at great peril. Thank you, Darling-Hammond, for forcing us to jump into the muddy waters of reality testing.

Richard Stiggins' chapter 6 "Measuring Performance in Teacher Assessment" is a helpful analysis of the role that performance assessment can play in the assessment of teachers and the steps that must be taken to insure the quality of those performance assessments. Performance assessment as he describes it has a philosophical and procedural kinship to Medley's chapter; perhaps equally true is that many of the concerns and cautions expressed in the Popham chapter might well be applied to what Stiggins is advocating. In a sense Stiggins is teaching us the ethics of performance assessment—the "thou shalls" and the "thou shall nots"—and in the process the criteria, decision points, and procedures are all systematically described.

It appears to this reader that more performance assessment, a higher quality of performance assessment, and more adequate training for performance assessors should all be instituted in the various settings for which Stiggins advocates performance assessment—particularly in teacher-education programs. His points are well taken and his advocacy is enlightened, especially if Popham's cautions are carefully considered and applied. My greatest concern is that in the process of applying the "shalls" and "shall nots" of performance assessment we may lose sight of the provisos that must be attached to those "shalls" and "shall nots" by the inevitable interactions and complexities of teaching. From Darling-Hammond's chapter we learn about the fearful lack of generalizability that seems to occur with teacher-effects research, and the consequences this has for teacher assessment. In Stiggins' chapter there seems to be a tendency to believe that performance assessment effectively applied will somehow overcome this complexity and lack of generalizability. In concentrating on the "shalls" and "shall nots" the interactions between the conceptual structure and the terms of the teaching milieu are not salient considerations, and this leads to some rather sweeping statements that may sometimes oversimplify the task at hand.

Two examples of this tendency appear in the discussion of the application of performance-assessment procedures to summative assessment. In discussing various decision contexts Stiggins indicates that "In each of these cases, the first requirement is that the

performance criteria be based on a thorough task analysis of the teaching process" (p. 205). Then in the next paragraph he reports that "The sample of exercises—whether naturally occurring or structured exercises—must reflect in a representative manner the full range of situations in which the student or teacher will be expected to demonstrate proficiency when teaching" (p. 205). A "thorough task analysis of the teaching process" and exercises that reflect the "full range of [teaching] situations" may be admirable ideals, but they may also be ideals that are contrary to the research evidence and not realizable in practice. Performance assessment is not a panacea that will solve all of the problems of teacher assessment or allow them to be ignored; it is a useful assessment procedure, with rules of application very well described by Stiggins, that is intended to add an additional measure of realism and validity to the assessment process. That realism and validity may be compromised if the complexity and interactions of the teaching situation are oversimplified or are not given the full attention required in both principle and practice.

Stiggins has provided an important service by warning us that in the hiring of teachers it is not always a "defensible assumption" that all preceding performance assessments were sound. The lack of training in performance assessment that may characterize many instructors, supervisors, and principals may be cause for justifiable skepticism. The effective use of performance assessment requires careful study of its concepts, principles, and rules of application. Stiggins' chapter provides a useful first resource for guiding that study.

Before commenting on the George Madaus chapter "Legal and Professional Issues in Teacher-Certification Testing: A Psychometric Snark Hunt," I should probably make a confessional statement about my predilections with respect to teacher-assessment tests. Madaus quotes a statement I made about one such teacher-assessment test, and that statement then reflected and doubtlessly continues to reflect my professional evaluations of most or all such teacher-assessment tests. My position on such matters is extremely similar to that of Madaus, and it is my considered judgment that Madaus is one of the most perceptive debunkers on the measurement scene since Oscar Buros passed away. If Madaus were not here, someone would have to invent him so that his clearheaded and realistic insights into what is really going on in teacher assessment would be available for all to ponder. Fortunately, he *is* here, and we can all profit immensely from his analysis and cautions.

From this preliminary statement I can then state without remorse the many Madaus contentions that I agree with, and also take up one issue that continues to perplex me in spite of my very extensive agreement. I agree with Madaus that "extant, generic multiple-choice teacher-certification tests make little sense, and are simply not valid" (p. 246). I agree that most or all validation studies generated by the commercial test publishers are "minimalist exercises designed to obtain a positive result" (p. 210), or at least usually achieve that result whether by conscious intent or not, and that current practices of content validation for these tests "redound to confirmation rather than disconfirmation" (p. 211). I agree with his statement that:

> the precondition of "legal defensibility" drives applied validation efforts to the detriment of a careful consideration of the evidence needed to sustain the inferences and decisions made from the test scores. The form and technique to construct a "legally defensible" test has almost completely overshadowed the essential question of the meaning behind the test score. (p. 226)

I also agree that teacher-assessment tests have been reified to such an extent that in the minds of some, particularly the public, they are perceived to measure that which even their designers did not design them to measure; I also wonder, however, whether those representing measurement have always done what they could to disabuse them of that notion. I particularly agree that content-validity evidence, based as it usually is on opinion alone, is not sufficient for teacher-certification tests generally and certainly not for the inferences typically drawn from them. I believe, along with Madaus, that the validation of teacher-certification tests must include evidence from all three traditional validity categories: content, criterion-related, and construct.

It is precisely at this point, however, that I begin to have qualms. It is not at all difficult to be consistently realistic in one's assessment of a very difficult problem and yet not be equally realistic in charting directions for its solution. In chapter 4 Mehrens tends to derogate the role that construct validity might have for licensure tests, and his thoughts about this exhibit a realistic tenor that should be considered as carefully as the realistic concerns brought up by Madaus with respect to the present status of teacher assessment as a whole. Yet Madaus advocates a "functional analysis of what minimally competent teachers actually do in their classrooms" (p. 246), for each and every area of certification, and the

generation of convincing criterion-related and construct-validity evidence appropriate thereto. That's a large order in view of the generally accepted conclusion that reliable and valid criteria are extremely difficult to identify in this area, and the generation of a nomological network of relationships is extremely hazardous because of the ill-defined nature of the construct of "teaching effectiveness" and the earlier described "shifting" criterion that can result both in teaching effectiveness attributable to entirely different causes and teaching effectiveness that interacts with student and situation. The comparison of correlations between what a given teacher-assessment test is supposed to measure and what it isn't supposed to measure may result in such small or nonexistent differences between the two (due to unreliable criteria and inherent construct definitional inadequacies) that defensible conclusions are difficult or impossible to draw. The construct of "teaching effectiveness" simply does not lend itself to construct-validity evidence as well as other constructs in psychology and education. In addition to the usual well-documented problems with establishing construct-validity evidence in the typical setting, we have in this instance a construct that is unusually difficult to work with in terms of construct-validity requirements.

Despite the tremendous difficulties in establishing criterion-related and construct-validity evidence for teacher-assessment instruments, I agree with George Madaus that it is absolutely essential that we try to do it. If we can do it, we are far ahead in the game; if we can't, perhaps that very fact can demonstrate that the present practice of relying on raw opinion euphemized as content-validity evidence must by comparison be even more hopelessly short of the goal. George Madaus has made an extremely important statement in this chapter; ignore it at extreme risk.

Ronald Berk's chapter "Limitations of Using Student-Achievement Data for Career-Ladder Promotions and Merit-Pay Decisions" is a well researched and very comprehensive account of how the public's most popular panacea for evaluating teachers can lead us into a morass of pitfalls. Although this is a more specialized chapter than most of the other chapters, dealing only with the use of student-achievement test scores to evaluate teachers, this kind of approach has such face validity for the public that it more than deserves this concentrated attention. It tends to evoke thoughts and concerns similar to those Jim Popham was discussing in his chapter warning about "Face Validity" as the "Siren Song for Teacher-Testers." Anything with this much face validity and this

many problems demands potent weaponry, and Berk certainly brings such weaponry to bear. First he reviews professional standards and court decisions relevant to his topic and concludes that "There are no professional standards or court decisions to support the use of achievement data for any type of teacher evaluation" (p. 298). Then he conducts a review of the research literature that identifies "several factors that can influence a teacher's measured effectiveness that are beyond his or her control" (p. 277); he identifies 42 such factors. To this he adds 11 factors relating to pretest-posttest gain that can cause confusion or make impossible any inferences regarding the teacher's actual contribution to such gain. He also discusses criteria for defining superior teacher performance and finds them wanting. It is a convincing exercise.

Regardless of the evidence that can be marshalled against the use of student-achievement–test gains as criteria of teaching effectiveness, however, the public is not likely to abandon this appealing gambit. We must return to Darling-Hammond's plea that we deal with the realities of our situation. Faced with the evidence that Berk has presented to us, we can attempt to accomplish at least three things: (a) we can take *every* opportunity to immerse the public in interpretable data and argument that will disabuse them of the notion that student-achievement test score gain is the panacea for evaluating teacher effectiveness; (b) we can make an equally strong effort to encourage the adoption of teacher-evaluation systems that make use of *multiple* indicators of teaching effectiveness in the hope that such systems will tend to counterbalance any constant measurement errors that may inhere in the individual indicators; and (c) we should forthrightly recognize that at this stage in our development any evaluation system (as mentioned earlier by this author) involves a value choice of those indicants that are perceived to define what we mean by teaching effectiveness, and in the absence of weightier logical and scientific evidence we should simply fill the gap with conscious choices consciously and openly defended. And we should always make clear exactly what we are doing, and the status of what we are doing in terms of knowledge base and scientific underpinnings (or lack thereof).

We turn now to a consideration of John Hoyle's chapter "Teaching Assessment: The Administrator's Perspective." When the symposium and this book were first planned, it was always a high priority to make sure that our treatment of teacher assessment

would make a very conscious and strong attempt to bridge the gap between theory and practice that often exists for such topics. With this in mind, we resolved to obtain some first-hand commentaries from representatives of two groups that are most integrally involved in the actual application of the teacher-assessment process: the administrators who administer the evaluation programs and often conduct the evaluations, and the teachers who are the objects of the evaluation and profit or suffer from them. To represent the administrator's perspective we selected John Hoyle of the Department of Educational Administration of Texas A&M University; to represent the teacher's perspective we selected Peg Shafer, who has been a teacher and who now represents teachers in her capacity as a teachers' union official. At the panel discussion I chaired at the conclusion of the symposium, I asked whether the researchers and the practitioners were speaking the same language when they discussed the topic of teacher assessment. There was a difference of opinion in response to that question; my personal opinion is that there is still an appreciable gap in communication and understanding. Readers may have felt that they were entering a different world as they read the Hoyle and Shafer contributions. Earlier we discussed the reality testing that Darling-Hammond required of us. Hoyle and Shafer force us into a reality-testing mode again, and with a vengeance. But the shock of reality testing is good for us, because it forces us to recognize that all the theorizing, researching, and discussion comes to naught unless it leads to practical outcomes that are sound in practice and facilitative of improvement.

In speaking from the administrator's perspective John Hoyle sounds the first jarring note when he asserts that the school principals who are required to do much of the teacher evaluation have only haphazard training at best in teacher evaluation, do not have the time to do an adequate job of teacher evaluation, and often feel there is a conflict between their evaluation and supervision roles. As a result the job typically doesn't get done very well. It seems that principals are not really obtaining a realistic grasp of what is going on in the classroom, either; it is interesting that he reports that one of the criticisms of the Texas Teacher Appraisal System was that "Teachers put on a good show when they are being observed because the criteria are so specific and fairly easy to follow" (p. 319). There is obviously a certain amount of game playing going on here, and it would be foolish not to recognize it as a source of invalidity in teacher evaluations. It also constitutes a good reason

for "involvement of the teachers in the entire developmental evaluation process" (p. 316), which is the first recommendation of the "One Best Model" format for a successful evaluation system.

The One Best Model system contains some recommendations that seem quite sound to me and may have some potential for reducing the gap between theory, research, and practice. They are worth quoting in their entirety:

> (a) involvement of the teachers in the entire developmental evaluation process, (b) performance criteria based on sound research and on local needs and concerns, (c) collaborative goal setting, (d) multidimensional methods for assessing teachers' skills, (e) careful analysis of data gathered in the assessment stage, (f) development of specific job targets, and (g) inclusion of a preobservation conference to acquire background data and a postobservation conference to mutually analyze classroom data and set goals for improvement. (p. 316)

There are certain features of these suggestions that recommend themselves because they effectively acknowledge the state of the art and the limits thereto. The involvement of teachers in the entire developmental evaluation process is good because it involves teachers in working with the system instead of against it. Performance criteria based on sound research and on local needs and concerns are good because they require a careful evaluation of available research and its practical utility, and the further recognition that value choice of criteria based on local needs and concerns is a perfectly legitimate practice if it is recognized for what it is—value choice. Collaborative goal setting is good because it again involves the teacher working with instead of against the system, and also emphasizes what may be by far the most important component of any evaluation system: the setting of goals for the future. Multidimensional methods for assessing teachers' skills are good because they implicitly acknowledge that any method or instrument is subject to its own peculiar error and criticism, and that there is greater safety in using multiple methods that can cast light on construct validity and counterbalance error. The development of specific job targets is good because it again focuses not on evaluation for evaluation's sake but rather on the all-important goals for future improvement. Maybe we can't always evaluate very well, but *any* evaluation that produces the outcome of a convinced teacher setting important professional goals has achieved the most important outcome of all.

One of the most interesting statements in the Hoyle chapter for

me is the statement that "any teacher evaluation form should include the following indicators:

- motivates students to achieve
- uses academic learning time effectively
- demonstrates proficiency in subject areas
- demonstrates command of the language
- promotes student academic growth
- learning objectives are clear
- learning strategies are based on objectives
- testing is based on objectives (p. 324)

The reason this statement was of such interest to me was that it seemed to illustrate several things about teacher-evaluation programs that are important to recognize. First, the list includes several items that almost no one would argue are not desirable for a teacher to exhibit. Second, in spite of the foregoing sentence many (perhaps the majority of) people would probably create a list of their own that would differ somewhat in terms of what was included and excluded and what the overall emphasis was. Third, there is nothing in this list that is mandated by research (including research on the prediction of student achievement), with the possible exception of time-on-task research. Fourth, when this list is examined in relation to the foregoing statements, the importance of *value choice* in defining teacher effectiveness emerges more potently than ever before. Many of the values represented are so general and generally accepted that they probably don't even stimulate much thought; the items they represent may be so general and vague that they are equally difficult to evaluate (e.g., "promotes student academic growth"). Other items are based on state-of-the-art ideology; the last three items, for example, all relate to *objectives;* yet I dare say you could identify teachers of excellence that would not rank high on these items and would yet deserve the label of *excellent* teacher. This again relates to the "shifting criterion" of teacher effectiveness mentioned earlier in this chapter; there may be 1,001 ways to be an effective teacher, and 1,001 ways to be an ineffective teacher. Into this mass of teacher behaviors you insert certain vectors that represent value-laden items, and if a given teacher has high loadings on the vectors you have inserted, that teacher is labeled a *superior, average,* or *inferior* teacher, as the case may be. But there is an interaction between the vectors chosen and teacher behavior such that one's standing could vary

from system to system, with repercussions of consequence. Notice what happens, for example, if I should enter the notion of "discipline" into the aforementioned list. In today's educational climate many would protest that my ideology was wrong, my values were wretched, my items were atavistic and inappropriate, and my intent was malicious. The maliciousness of my intent, however, is limited only to illustrating the importance of *value choice* in the defining of teacher effectiveness and its measurement. In the absence of definitive research, value choices, pure and simple, step in to fill the gap.

Peg Shafer's chapter "Appraisal: The Teacher's Perspective" may have impressed some readers as being comprehensively negative about most or all forms of teacher appraisal. Peg Shafer's presence in person at the symposium, however, projected a somewhat different image from Peg Shafer's words in writing. Her presentation was characterized by cordiality, spontaneity, vigor, and a deep commitment to realism and honesty. Here is a person who is very deeply identified with teachers and their aspirations, needs, and problems, and she intends to "tell it like it is" and set the record straight. As a well-respected and effective leader of teachers she provides us with a striking opportunity to perceive appraisal issues as those do who are most affected by them. When we planned this symposium and book, we decided it was important to have this point of view, and we got it in full measure. Anyone who ignores teacher reactions as Peg Shafer describes them might as well give up all hope of effecting a valid teacher-evaluation system. As she says in her own inimitable manner, "If teachers believe in their hearts that appraisal is a farce, they will scuttle the plan in their churches, the grocery stores, their neighborhoods, and everywhere they feel safe to speak their minds" (p. 341). That's not advice to be taken lightly.

There is a great deal said in this chapter, and it is impossible for me to comment on it all; instead I shall briefly discuss some issues that particularly piqued my interest as I read the chapter. One point I may have known before but perhaps did not recognize for its widespread implications is that teachers often perceive teacher-evaluation programs as public relations gimmicks. It also appears that the more "farcical" the evaluation program is in their eyes, the more it is perceived as PR. It would be wrong to dismiss this as mere cynicism, because this attitude undoubtedly contains at least some measure of truth. Principals may also have doubts about the validity or utility of some teacher-evaluation programs, but they

go through the motions because the motions represent a response, whether adequate or not, to the public's cry for getting rid of the incompetents and identifying superior teachers for reward. Public dissatisfaction with schools and teachers must somehow be responded to, and a teacher-evaluation system may be regarded as a good first line of defense, or at least one of them. What all this underscores is that a teacher-evaluation system must be perceived as having some inherent real benefits for teachers in order to compensate for an apparently ubiquitous and strong tendency for them to perceive such systems as mere PR. If it is not realistic to assume that teacher-evaluation systems can have real benefits for teachers, then passive-aggressive behavior will be the likely result and the "selling" job will be correspondingly difficult. Teacher-evaluation programs perceived by teachers and others as mostly PR mechanisms define a setting ripe for disgruntlement and low morale.

Another issue that had tremendous import for me was well expressed by two excerpts from the Shafer presentation:

> From the belief that a model of good teaching exists, the public can take the short step to the harmful conclusion that anyone can teach if they possess a passable knowledge of a subject and are able to emulate the behaviors in the model. They can erroneously conclude that any deviation from the model must be faulty. . . . (p. 333).
>
> The worst news I have for you today is that so far, teachers have not bought into the recent changes in teacher evaluation. Teachers sense danger and they have circled their wagons. The clearest danger is that researchers and testing experts are searching for ways to reduce teaching to paper so that we can convince the public that we can control and improve the way teachers perform. But teachers know that nothing unique and exceptional ever grew from a dry formula. Creativity and flair can't be standardized. Painting by number has never produced any masterpieces. Resistance to "recipes" for improvement is growing every day, and there is massive resistance to standardized methods of teaching, both from individuals and organizations. (pp. 336–337)

These two passages are reminiscent of the "shifting criterion" I have occasionally referred to as constituting a problem for teacher evaluation and my conjecture, colloquially expressed, that there may be 1,001 ways to be an effective teacher and 1,001 ways to be an ineffective teacher. You cannot fault teachers for concluding that teaching is an art rather than a science; many of them have concluded that the science we have offered them is conflicting and

inconclusive and of little help. To the extent that any teacher-evaluation system presents or implies a rather singular model for effective teaching, they may have a right to rebel. As a university teacher I would rebel myself. Shafer is right to say that "Creativity and flair can't be standardized." But it can be thwarted by evaluation systems that rigidly define by competencies or otherwise what an effective teacher is, and then leave little room or acknowledgment for departures from the standard ideal that may be as effective or more effective than what was originally defined. If teacher-evaluation systems turn out to be procrustean beds fashioned by people who claim to know, on insufficient evidence, what "the" effective teacher is, the rank-and-file teacher has a right to be disbelieving and defensive.

Many people who read Shafer's chapter may conclude that she and the teachers she represents are being *too* defensive. They may be right. But the concerns that Shafer describes must be carefully attended to for purposes of good communication. One of the often-stated standard requirements of an effective teacher-evaluation program is that teachers should be an integral part of the developmental process for such a program. If that is so, teachers should be heard, and heard well. Shafer has helped us to hear them well.

This completes my discussion of the contributions made by chapter authors to this book on teacher assessment. These authors have addressed varied topics and have certainly provided a wide variety of professional judgments and opinions. My own comments have also been wide ranging; as a former dean of a college of education, a vice president for academic affairs, director of the Buros Institute of Mental Measurements, and now a university professor again, I have had the benefit of many different perspectives, and these different perspectives undoubtedly came into play as I reacted to a topic as encompassing and with as many ramifications as this one. Much work still remains to be done in this area, and we have tried to consider some of the questions to be addressed for further progress to occur. The title of this chapter makes reference to concluding thoughts and lingering questions about this topic of teacher assessment. If I were to summarize these concluding thoughts and lingering questions with a few broad statements, those statements would include the following:

1. There is great danger in oversimplifying the task of teacher assessment, or allowing the hazardous oversimplification of the teacher assessment task by the public.

2. The appeal to face validity in teacher-assessment tasks may bring satisfaction in the short term but invalid inferences as a final outcome.

3. Any teacher-evaluation system must allow for multiple "correct" responses that do justice to the complexity of teaching and acknowledge that there are many *patterns* of teaching behavior that can be designated as "effective" teaching.

4. The validation of teacher-assessment instruments and procedures is fraught with many difficulties, and there is a difference of opinion among measurement professionals about what kinds of evidence are acceptable and required.

5. The determination of cut scores and the validity of cut-score decisions continue to be major issues with most teacher-assessment tests.

6. Many factors other than teacher behaviors have profound effects on student learning, and any teacher-assessment system that does not take this into account is closing it eyes to reality and confounding its efforts to predict.

7. The legal, political, social, and organizational context of teacher-assessment efforts has immense impact on what is developed and the nature of its consequences.

8. In the absence of definitive scientific findings, value choice plays a significant role in the selection of variables for the development and implementation of teacher assessment programs.

9. The attitudes of teachers and principals to a teacher-assessment program can make a tremendous difference in its acceptance, validity, and consequences; teachers are truly the "make or break" agents in the implementation of a teacher-assessment program.

10. Clear communication among the many parties interested in teacher assessment is critical; at the present time such communication is in short supply, and both teacher assessment and the democratic process suffer as a result.

I hope the present volume has helped to meet the critical need expressed in this last statement for clarification of issues and facilitation of communication.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educa-*

tional and psychological testing. Washington, DC: American Psychological Association.

Bush, R. N. (1979). Implications of the BTES. *The Generator, 9*(1), 13–15.

Dewey, J. (1949). *Experience and education.* New York: Macmillan.

Doyle, W. (1979). Classroom tasks and students' abilities. In P. L. Peterson & H. J. Walberg (Eds.), *Research on teaching* (pp. 183–209). Berkeley, CA: McCutchan.

Knapp, M. S. (1982). *Toward the study of teacher evaluation as an organizational process: A review of current research and practice.* Menlo Park, CA: Educational and Human Services Research Center, SRI International.

Piper, M. K., & O'Sullivan, P. S. (1981). The National Teacher Examination: Can it predict classroom performance? *Phi Delta Kappan, 62*(5), 401.

Quirk, T. J., Witten, B. J., & Weinberg, S. F. (1973). Review of studies of the concurrent and predictive validity of the National Teacher Examinations. *Review of Educational Research, 43*(1), 89–113.

Rudner, L. M. (1988). Teacher testing—An update. *Educational Measurement: Issues and Practice, 7*(1), 16–19.

Task Force on Teaching as a Profession. (1986). *A nation prepared: Teachers for the 21st century.* Hyattsville, MD: Carnegie Forum on Education and the Economy.

Webb, L. D. (1983). Teacher evaluation. In S. B. Thomas, N. H. Cambron-McCabe, & M. M. McCarthy (Eds.), *Educators and the Law* (pp. 69–80). Elmont, NY: Institute for School Law and Finance.